Regime Change: Bit-Depth versus Measurement-Rate in Compressive Sensing*

Jason N. Laska and Richard G. Baraniuk†

October 18, 2011

Abstract

The recently introduced *compressive sensing* (CS) framework enables digital signal acquisition systems to take advantage of signal structures beyond bandlimitedness. Indeed, the number of CS measurements required for stable reconstruction is closer to the order of the signal complexity than the Nyquist rate. To date, the CS theory has focused on real-valued measurements, but in practice, measurements are mapped to bits from a finite alphabet. Moreover, in many potential applications the total number of measurement bits is constrained, which suggests a tradeoff between the number of measurements and the number of bits per measurement. We study this situation in this paper and show that there exist two distinct regimes of operation that correspond to high/low signal-to-noise ratio (SNR). In the *measurement compression* (MC) regime, a high SNR favors acquiring fewer measurements with more bits per measurement; in the *quantization compression* (QC) regime, a low SNR favors acquiring more measurements with fewer bits per measurement. A surprise from our analysis and experiments is that in many practical applications it is better to operate in the QC regime, even acquiring as few as 1 bit per measurement.

1 Introduction

The *compressive sensing* (CS) framework has sparked renewed interest in sampling and signal acquisition [1, 2]. The framework can be concisely summarized by three fundamental components: *i*) *underdetermined linear measurement systems*, i.e., we obtain the measurements

$$\boldsymbol{y} = \Phi \boldsymbol{x} + \boldsymbol{e}, \tag{1}$$

of the signal $\boldsymbol{x} \in \mathbb{R}^N$, with $\Phi \in \mathbb{R}^{M \times N}$ and $M \ll N$, and with measurement error $\boldsymbol{e} \in \mathbb{R}^M$; *ii*) *structured signal models*, such as K-sparse signals, i.e., $\boldsymbol{x} \in \Sigma_K := \{\boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{x}\|_0 := |\mathrm{supp}(\boldsymbol{x})| \leq$

*This work was supported by the grants NSF CCF-0431150, CCF-0728867, CCF-0926127, CNS-0435425, and CNS-0520280, DARPA/ONR N66001-08-1-2065, N66001-11-1-4090, ONR N00014-07-1-0936, N00014-08-1-1067, N00014-08-1-1112, and N00014-08-1-1066, AFOSR FA9550-07-1-0301 and FA9550-09-1-0432, ARO MURI W911NF-07-1-0185 and W911NF-09-1-0383, and the Texas Instruments Leadership University Program.

†Department of Electrical and Computer Engineering, Rice University, Houston, TX, 70015 USA. Email: laska@rice.edu, richb@rice.edu.

$K\}$; and *iii*) *computational reconstruction*, one example being the convex program known as *Basis Pursuit Denoising* (BPDN),

$$\widehat{\boldsymbol{x}} = \min_{\boldsymbol{x}\in\mathbb{R}^N} \|\boldsymbol{x}\|_1 \;\text{ s.t. }\; \|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2 < \epsilon, \tag{2}$$

that guarantees $\|\boldsymbol{x}-\widehat{\boldsymbol{x}}\|_2 \leq C\epsilon$ for $\|\boldsymbol{e}\|_2 < \epsilon$, C a constant, and under certain conditions on Φ [3]. A significant body of work has been devoted to the study of each of these components individually, e.g., by *a*) characterizing conditions on Φ that provide robust mappings of sparse signals and designing physical sampling systems that satisfy such conditions [3–8]; *b*) proposing more refined classes of highly structured signals [9–11]; and *c*) providing reconstruction guarantees and fast solvers for BPDN and other convex programs [12–15] as well as greedy and first-order algorithms [16–18].

CS promises to lessen our sampling burden. The simple consequence of (1) is that, when the acquisition of each measurement is "expensive," we benefit by sensing only M values rather than N. One example of such a situation is magnetic resonance imaging (MRI) [19]. We seek to minimize the amount of time to image a patient; however, each measurement is time-consuming, leading to a total acquisition time that is currently on the order of tens of minutes. Another example is sampling, where the required Nyquist rate for wideband signals may be prohibitively highy [5, 20]. It is possible to design a physical sampling system $\bar{\Phi}$ such that $\boldsymbol{y} = \Phi\boldsymbol{x} = \bar{\Phi}(x(t))$ where $\boldsymbol{x}$ is a vector of Nyquist-rate samples of a bandlimited signal $x(t)$, $t \in \mathbb{R}$. In this case, (1) translates to low, sub-Nyquist sampling rates, a potential boon for wideband acquisition.

In practice, three issues may arise during signal acquisition that are not modeled by (1). First, the real-valued CS measurements will be mapped to discrete bits via a quantizer. Second, there may be noise present on the input signal. Third, we often must limit the total number of measured bits $\mathfrak{B}$, i.e., we are constrained by a bit-budget when transmitting or storing the measurements. Thus, a more precise model of CS acquisition is

$$\boldsymbol{y}_Q = \mathcal{Q}_B(\Phi(\boldsymbol{x} + \boldsymbol{n}) + \boldsymbol{e}), \tag{3}$$

where the *signal noise* is denoted by $\boldsymbol{n} \in \mathbb{R}^N$, and $\mathcal{Q}_B : \mathbb{R} \to \mathfrak{A}$ is a B-bit scalar quantization function (applied element-wise in (3)) that maps real-valued CS measurements to the discrete alphabet $\mathfrak{A}$ with $|\mathfrak{A}| = 2^B$. In this paper we will model $\boldsymbol{n}$ as a random vector with each element having variance $\sigma^2_{\boldsymbol{n}}$. Since the primary source of measurement noise in a well-designed hardware system derives from quantization, we will assume $\|\boldsymbol{e}\|_2 = 0$.[1] Since the quantizer is scalar, we can write the bit-budget constraint as

$$\mathfrak{B} = MB. \tag{4}$$

Although we will focus on scalar quantization in this paper, alternative quantization techniques such as sigma-delta [21] or non-monotonic scalar quantization [22] have also been proposed for CS systems, as have many algorithms specialized to CS quantization problems [23–29]. The main themes presented here will be generally applicable to these techniques and algorithms as well.

The fixed bit-budget $\mathfrak{B} = MB$ and the signal noise $\boldsymbol{n}$ impose a competing performance tradeoff as a function of M. On the one hand, since $B = \mathfrak{B}/M$, we can increase the bit-depth as we decrease the number of measurements, thereby increasing the precision of each measurement. On the other hand, signal noise is amplified due to *noise folding* as we decrease the number of measurements,

[1] The general trends presented in this paper remain unchanged when $\|e\|_2 > 0$.

thereby decreasing the precision of each measurement [30].[2] Thus, we find ourselves in somewhat of a conundrum: as we take fewer measurements we can allocate more bits per measurement (good), but noise folding increases the risk of wasting these bits on already imprecise measurements (bad).

We can gain more insight into this conundrum through a back-of-the-envelope calculation of the optimal total acquisition error, which comprises the expected mean-squared distortion due to a scalar quantizer for Gaussian measurements $O(\|x\|_2^2 2^{-2B})$ and the expected reconstruction error due to measurement noise $O\left(\frac{N}{M}\sigma_{\boldsymbol{n}}^2\right)$. Equating these noise levels to minimize the total mean square error (MSE) leads to

$$B \approx \frac{1}{2}\log_2\left(\frac{\|\boldsymbol{x}\|_2^2}{\sigma_{\boldsymbol{n}}^2}\frac{M}{N}\right).$$

This expression can also be found using classical rate-distortion bounds in terms of the signal-to-noise ratio (SNR) [33, 34]. Imposing the fixed bit-budget $B = \mathfrak{B}/M$ and rearranging terms, we find that the MSE is minimized when

$$\log_2\left(\frac{\|\boldsymbol{x}\|_2^2}{N\sigma_{\boldsymbol{n}}^2}\right) \approx \frac{2\mathfrak{B}}{M} - \log_2(M). \tag{5}$$

The term on the left is the logarithm of the SNR of the input signal. For fixed $\mathfrak{B}$ and N, (5) implies that there are two operational regimes that correspond roughly to "high" input SNR and "low" input SNR. At high input SNR, the MSE is minimized by taking a small number of measurements M with large bit-depth; we call this the *measurement compression* (MC) regime. At low input SNR, the MSE is minimized by taking a large number of measurements M with small bit-depth; we call this the *quantization compression* (QC) regime. The exact SNR at which the transition between the two regimes occurs is a function of the total bit-budget. A primary contribution of this paper is to expose and explore the QC regime.

In this paper we argue for the distinction between the MC and QC regimes in two ways. First, we formalize the back-of-the-envelope calculation in (5) by analyzing the reconstruction MSE that results from the combined effects of quantization and signal noise folding. Specifically we provide an upper bound on this MSE for an optimal non-uniform scalar quantizer that roughly predicts the trends of the optimal bit-depth for different signal noise powers and bit-budgets. Second, we provide a suite of simulations for a specific setup frequently encountered in practice: the acquisition of sparse signals from uniformly quantized measurements. Surprisingly, at certain practical SNRs, our simulations suggest that a 1-bit quantizer (using the reconstruction techniques developed in [35]) exhibits better performance than larger bit-depth quantizers.

Revisiting the example CS applications from above, a CS MRI device should aim to operate in the MC regime, since the total data acquisition time is proportional to M. In this case, (5) recommends acquiring high SNR measurements and quantizing them finely. In contrast, a low SNR wideband sampling system should aim to operate in the QC regime. In this case, (5) recommends acquiring low SNR measurements and quantizing them coarsely. Fortunately, by some divine Providence, sampling rate and bit-depth enjoy an inverse relationship in practical ADCs; specifically, we obtain an exponential increase in sampling rate as the bit-depth is decreased [36]. Taking this idea to its logical extreme, it has been shown that it is possible to drive the bit-depth down to 1 bit per CS measurement and still guarantee stable signal recovery [35, 37–39]. In this case the quantizer is simply a comparator, enabling an extremely high sampling rate.

[2] Roughly speaking, noise folding implies that during reconstruction we lose about 3dB of signal-to-noise ratio (SNR) as the number of measurements is halved [31, 32].

The remainder of this paper is organized as follows. In Section 2, we provide the necessary CS background for our analysis and simulations. In Section 3, we develop a bound on the reconstruction error due to quantization and signal noise, expressed in terms of a fixed bit-budget. In Section 4, we present a series of numerical simulations that further support our argument. We conclude in Section 5 with a discussion on the implications of this work.

2 Background

2.1 CS Toolkit

Before examining the effect of noise and quantization on CS reconstruction performance, we first review a few key results and definitions that enable our analysis.

CS reconstruction can be interpreted as consisting of two steps: first finding the non-zero coefficient locations (the support) and then estimating the coefficient values. If we can correctly identify the true signal support, then the optimal linear estimate for coefficient values can be computed via least squares:

$$\widehat{\boldsymbol{x}}|_{\Omega} = \Phi_{\Omega}^{\dagger}\boldsymbol{y}, \quad \widehat{\boldsymbol{x}}|_{\Omega^C} = \boldsymbol{0}, \tag{6}$$

where Φ_Ω denotes the submatrix of Φ formed by selecting the columns of Φ according to the index set Ω, $\widehat{\boldsymbol{x}}|_\Omega$ is the corresponding subvector of $\widehat{x}$, Ω^C is the complement set to Ω, and $\dagger$ denotes the Moore-Penrose pseudo-inverse. Indeed, if an oracle were to provide the true support Ω, then no linear CS reconstruction algorithm can perform better than (6). Thus, reconstruction with known signal support is sometimes called *oracle-assisted* reconstruction [32, 40]. Our analysis will be primarily in terms of the performance of this best-case reconstruction algorithm. Furthermore, from [22, 35], when there is no noise on the measurements, the reconstruction (6) is also consistent, meaning that

$$\mathcal{Q}_B(\Phi\hat{\boldsymbol{x}}) = \mathcal{Q}_B(\Phi_\Omega\hat{\boldsymbol{x}}|_\Omega) = \mathcal{Q}_B(\Phi_\Omega\Phi_\Omega^{\dagger}\boldsymbol{y}) = \mathcal{Q}(\boldsymbol{y}) = \boldsymbol{y}.$$

There is no better nonlinear estimator for the quantized measurements than a consistent estimator.

Robust reconstruction guarantees will only hold for measurement systems Φ that are "well-conditioned." For instance, the so-called *restricted isometry property* (RIP) of a matrix Φ has been shown to be a sufficient condition for the robust recovery of sparse signals via several algorithms [3, 16]. The RIP of order K with constant δ is defined as

$$(1-\delta)\|\boldsymbol{x}\|_2^2 \leq \|\Phi\boldsymbol{x}\|_2^2 \leq (1+\delta)\|\boldsymbol{x}\|_2^2, \tag{7}$$

for all $\boldsymbol{x} \in \Sigma_K$. Roughly speaking the RIP ensures that the norm of the measurements is close to the norm of the signal for all K-sparse signals. An alternative way of thinking of this is that the singular values of any submatrix formed by K or fewer columns of Φ are bounded close to 1; hence any K-column submatrix of Φ is close to an isometry.

The RIP ensures stable oracle-assisted recovery when white noise is added to the measurements. Specifically, suppose that $\boldsymbol{z} = \Phi\boldsymbol{x} - \boldsymbol{y}$, where $\boldsymbol{z}$ is a zero-mean random vector with uncorrelated (white) entries, each having variance $\sigma_{\boldsymbol{z}}^2$. Furthermore suppose that Φ has the RIP of order K, and that $\boldsymbol{x}$ is K-sparse. Then Theorem 4.1 of [32] demonstrates that oracle-assisted reconstruction will have expected error

$$\frac{K\sigma_{\boldsymbol{z}}^2}{1+\delta} \leq \mathbb{E}(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2^2) \leq \frac{K\sigma_{\boldsymbol{z}}^2}{1-\delta}. \tag{8}$$

A key component of our analysis below will be understanding the variance of the noise term $\boldsymbol{z}$ that arises from the quantized noisy measurements $\boldsymbol{y}_Q$. The expression (8) then gives the intuition that the expected reconstruction error behaves on the order of the variance of the error per measurement $\sigma_{\boldsymbol{z}}^2$.

We will also make use of a result that relates the variance $\sigma_{\boldsymbol{n}}^2$ of the signal noise to the variance of the measured noise $\sigma_{\Phi\boldsymbol{n}}^2$. If $\boldsymbol{n}$ is white with mean zero and variance $\sigma_{\boldsymbol{n}}^2$, and Φ has orthonormal rows, i.e., $\Phi\Phi^T = \frac{N}{M}\boldsymbol{I}_M$,[3] then it is straightforward to show that the measured noise is also white and zero mean and has variance

$$\sigma_{\Phi\boldsymbol{n}}^2 = \frac{N}{M}\sigma_{\boldsymbol{n}}^2. \tag{9}$$

Note that the measured noise is only uncorrelated (i.e., white) when $M \leq N$; indeed, the condition $\Phi\Phi^T = \frac{N}{M}\boldsymbol{I}_M$ can only hold when $M \leq N$.

In [32], the authors combine the results of (8) and (9) to obtain a bound on the oracle-assisted reconstruction error due to noise folding. We will take a similar approach, however we will additionally include the effects of quantization. Furthermore, because our quantization error is not necessarily uncorrelated, we first generalize (8) to obtain an upper bound on the oracle reconstruction error with uncorrelated measurement noise.

2.2 1-bit CS

The results of the conventional CS framework above will enable us to analyze scalar quantized measurements when the bit-depth is greater than 1. However, CS measurements can be coarsely quantized to just 1 bit, representing their signs. These facts preclude 1-bit CS from being analyzed within the conventional linear CS framework. Even though meaningful theoretical comparisons are difficult to make between 1-bit and conventional CS, it is beneficial to compare their empirical performances, since both types of CS can be useful in practice. Thus, we very briefly review the key results of the 1-bit CS framework [35, 37]. Formally, 1-bit measurements can be written as

$$\boldsymbol{y}_s = A(\boldsymbol{x}) := \operatorname{sign}(\Phi\boldsymbol{x}). \tag{10}$$

To reconstruct, we search for a sparse, unit-norm signal $\hat{\boldsymbol{x}}$ that is *consistent* with the measurements, meaning that $A(\hat{\boldsymbol{x}}) = A(\boldsymbol{x})$. We restrict our attention to unit-norm signals, since the scale of the signal is lost during the 1-bit quantization process. This problem is generally non-convex, and thus it is difficult to design an algorithm that will be guaranteed to find the desired solution. Nonetheless several algorithms have been proposed to approximately solve this problem [35, 37–39]; convex programs have also been formulated [41].

In much the same way that the RIP of Φ guarantees stable reconstruction from ℓ_1-minimization programs [42], the so-called *binary ϵ-stable embedding* (BϵSE) provides a similar robustness for the mapping A with consistent algorithms [35]. The property explains that the normalized Hamming distance between any two sets of measurements is within ϵ of the normalized angular distance between the original signals, for all unit-norm K-sparse signals. It can be shown that, if the elements of Φ are drawn from a Gaussian distribution, then Φ satisfies the BϵSE with high probability, and thus CS systems that enable 1-bit quantized measurements exist.

[3] The so-called *tight frame* condition $\Phi\Phi^T = \frac{N}{M}\boldsymbol{I}_M$ is not overly restrictive, since for any RIP matrix Γ, a matrix that has both the same row-space as Γ and the tight frame condition can be derived from Γ [32].

Our simulations will make use of two 1-bit CS algorithms originally introduced in [35]. Specifically, we will employ the BIHT and BIHT-ℓ_2 algorithms. The former can be thought of as minimizing a one-sided ℓ_1-norm and imposing a sparse unit-norm signal model, while the latter can be thought of as minimizing a one-sided ℓ_2-norm instead. By one-sided norm, we mean that the positive elements of a vector are set to zero before the norm is computed. The BIHT algorithm has been shown to perform better in low noise scenarios, while the BIHT-ℓ_2 algorithm has been shown to perform better in high noise scenarios [35].

3 Analysis of Quantized CS Systems with Signal Noise

In this section we derive a new upper bound on the oracle-assisted reconstruction error due to both noise and quantization, making the back of the envelope calculation (5) more rigorous. This bound enables us to argue that, for a fixed bit-budget $\mathfrak{B} = MB$, it may be better to quantize to fewer bits per measurement B than take fewer measurements M. The following theorem is proved in Appendix A.

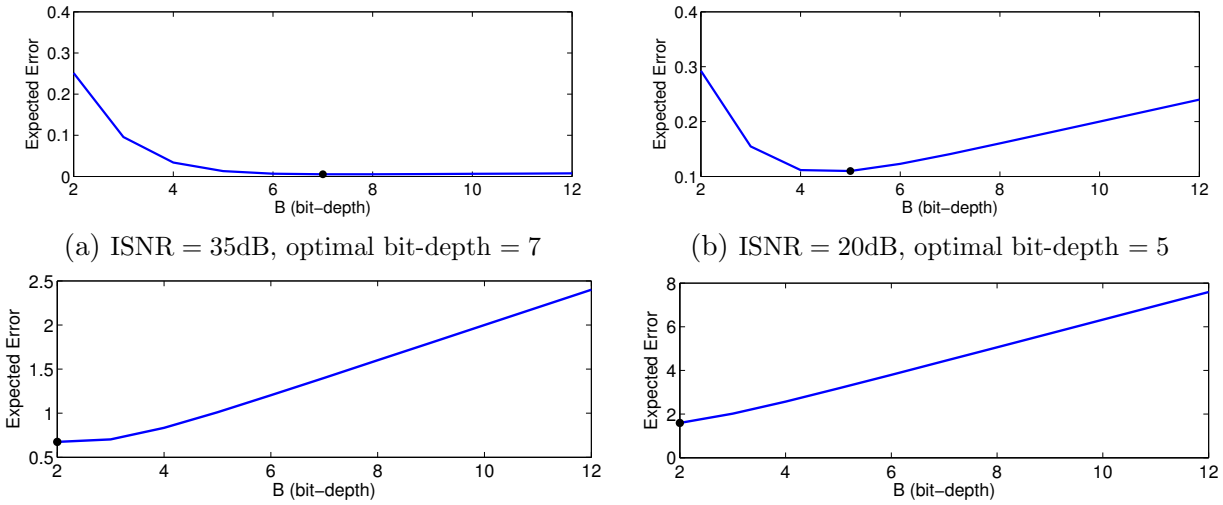
Theorem 1. *Suppose that $\boldsymbol{y}_Q = \mathcal{Q}_B(\Phi(\boldsymbol{x} + \boldsymbol{n}))$. Let the signal $\boldsymbol{x} \in \mathbb{R}^N$ be sparse with support $\Omega \in \{1, \dots, N\}$ and $|\Omega| = K$, where the elements Ω are chosen uniformly at random and the amplitudes of the non-zero coefficients are drawn according to $x_j \in \Omega \sim \mathcal{N}(0, \sigma_{\boldsymbol{x}}^2)$. Let the signal noise $\boldsymbol{n} \in \mathbb{R}^M$ be a random, white, zero-mean vector with variance $\sigma_{\boldsymbol{n}}^2$. Furthermore, let the $M \times N$ matrix Φ satisfy the RIP of order K with constant δ, $\Phi\Phi^T = \frac{N}{M}\boldsymbol{I}_M$, and $M < N$. Choose $\mathcal{Q}_B$ to be the optimal scalar quantizer with $B > 1$ that minimizes the MSE for the distribution of the measurements $\Phi(\boldsymbol{x} + \boldsymbol{n})$. Then for a fixed bit-budget of $\mathfrak{B} = MB$, the MSE of the oracle-assisted reconstruction estimate $\widehat{\boldsymbol{x}}$ satisfies*

$$\mathbb{E}\left(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2^2\right) \leq \frac{2K}{\mathfrak{B}(1-\delta)}\left(K\sigma_{\boldsymbol{x}}^2 B 2^{-2B} + N\sigma_{\boldsymbol{n}}^2 B\left(1 + 2^{-2B}\right)\right) + \frac{K}{(1-\delta)}\left(\frac{\mathfrak{B}}{B} - 1\right)\mathfrak{S}, \tag{11}$$

where $\mathfrak{S} = \max_{i \neq j} |\mathbb{E}(\mathcal{Q}_B(\Phi\boldsymbol{x} + \Phi\boldsymbol{n})_i \mathcal{Q}_B(\Phi\boldsymbol{x} + \Phi\boldsymbol{n})_j)|$ is the correlation between the quantized measurements.

Each component of the bound (11) is fairly intuitive. The term $K\sigma_{\boldsymbol{x}}^2 B 2^{-2B}$ reflects the error due to quantizing the measurements. The term $N\sigma_{\boldsymbol{n}}^2 B\left(2^{-2B} + 1\right)$ reflects both the error due to measured signal noise as well as the quantization of that noise. The reconstruction error is effectively proportional to these two terms. The final term $\left(\frac{\mathfrak{B}}{B} - 1\right)\mathfrak{S}$ reflects an additional error due to the correlation between the quantized measurements. In many CS scenarios we expect this term to be close to zero, and furthermore for large B it has been shown that this term can be accurately approximated as zero [43]. Thus, choosing the optimal B primarily comes down to balancing the terms inside the parentheses.

The bound in (11) applies to strictly sparse signals immersed in signal noise. However, it may also be of interest to consider so-called *compressible signals*, i.e., signals that are not strictly sparse but that can be reasonably approximated by retaining their K largest magnitude coefficients. For such signals, the "tail" part of the signal that we do no expect to recover, i.e., the subset of the smallest $N - K$ entries, is also subject to noise folding. Theorem 1 can be extended to handle compressible signals by inflating the second term to account for the additional correlation between the quantized measurements. The general performance trends will be similar to sparse signals in



(a) ISNR = 35dB, optimal bit-depth = 7

(b) ISNR = 20dB, optimal bit-depth = 5

(c) ISNR = 10dB, optimal bit-depth = 2

(d) ISNR = 5dB, optimal bit-depth = 2

Figure 1: Upper bound on the oracle-assisted reconstruction error as a function of bit-depth B and ISNR. The term inside the parenthesis in the bound (11) was computed. Black dots denote the minimum point on each curve.

noise; i.e., signals that are "less compressible" will induce the same regime as signals with low input SNR.

The bound in (11) is pessimistic, since we do not take into account the benefits accrued by increasing the number of measurements, for instance by improving the RIP constants of Φ. Furthermore, when the quantization error is large enough to dominate the measurement noise, the measurement noise terms may not play an active role in the true behavior of the system. Again, this is not reflected by the bound. Finally, the bound does not apply to 1-bit quantization or the case where $M > N$.

To use the bound (11) to support our argument that there are both MC and QC regimes in CS, we examine the behavior of the oracle-assisted reconstruction error as a function of the bit-depth B (or equivalently the number of measurements M since $\mathfrak{B} = MB$). Since the solution for the optimal B cannot be computed in closed form without resorting to tabulated functions, we evaluate the bound over some interesting parameters. The evaluation of the bound is depicted in Figure 1, where plots (a)–(d) correspond to input signal-to-noise ratios (ISNRs) of 35dB, 20dB, 10dB, and 5dB, respectively. We define the *input SNR* (ISNR) in dB as

$$\text{ISNR} := 10 \log_{10} \left(\frac{\mathbb{E}(\|\boldsymbol{x}\|_2^2)}{\mathbb{E}(\|\boldsymbol{n}\|_2^2)} \right). \tag{12}$$

where $\mathbb{E}(\|\boldsymbol{x}\|_2^2) = K\sigma_{\boldsymbol{x}}^2$ and $\mathbb{E}(\|\boldsymbol{n}\|_2^2) = N\sigma_{\boldsymbol{n}}^2$.

Since we are primarily concerned with the performance trend of (11) as a function of B and the ISNR, we make a few simplifications when plotting the bound. First, we only evaluate the term inside the parenthesis; this term is proportional to the error on the measurements and does not depend on the RIP constant, the sparsity K, or the correlation between the quantization errors. Second, by only evaluating the term inside the parenthesis in (11), we do not take into account the effect of M on the RIP constants (δ decreases as M increases). The minimum error point in each curve is denoted by a solid black dot.

The message from Figure 1 is clear. The tradeoff between the number of measurements M and bit-depth B empirically follows a convex curve, i.e., the error not only increases when B is too small, but the error also increases when B is too large. In other words, more bits per measurement is not necessarily optimal. Furthermore, as expected, the minimum reconstruction error occurs for smaller B as the ISNR decreases. For the high ISNR of 35dB, the bound is minimized at a bit-depth of approximately 7 bits per measurement. The is an example of the MC regime, where larger bit-depths and thus lower M yield the best performance. For the low ISNR of 10dB, the bound is minimized at a bit-depth of approximately 2 bits per measurement. This is an example of the QC regime, where larger bit-depths and thus higher M yield the best performance.

4 Experiments

In the previous section we have argued that the QC regime exists by deriving an upper bound on the oracle-assisted reconstruction error. In this section we perform a suite of simulations to empirically study for which input noise levels and bit-budgets this regime will occur in practical systems. Specifically our simulations *i*) validate the theoretical result in Theorem 1, *ii*) demonstrate the performance achieved in practice when combining quantization and signal noise, and finally *iii*) prove the existence of the QC regime. A surprising additional result emerges from the simulations: when nontrivial signal noise is present, 1-bit CS systems perform competitively with, if not better than conventional CS with uniform multibit quantization.

4.1 Setup

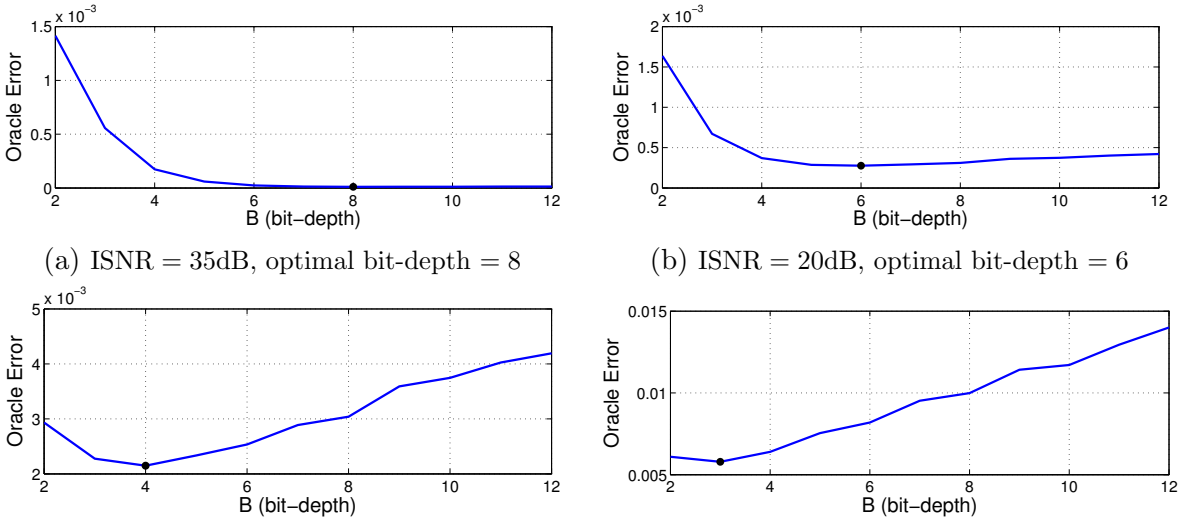
Our simulations were performed using canonically (identity) sparse signals $\boldsymbol{x}$.[4] The signals were measured with i.i.d. Gaussian matrices, i.e., $\boldsymbol{y} = \Phi(\boldsymbol{x}+\boldsymbol{n})$ where the matrix Φ has elements $\phi_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/M)$. The measurements were quantized uniformly with quantization interval $\Delta = T2^{-B+1}$, where T is the dynamic range of the quantizer. In all simulations, we chose $T = \|\Phi\boldsymbol{x}\|_\infty$ to maximize the range of the quantizer and ensure that for any noiseless measurement $|(\Phi\boldsymbol{x})_i - \mathcal{Q}_B((\Phi\boldsymbol{x})_i)| \leq \Delta/2$.

In each trial we drew a new $M \times N$ sensing matrix Φ and a new signal $\boldsymbol{x}$. The non-zero coefficients of $\boldsymbol{x}$ were chosen according to a Gaussian distribution, and their positions were chosen at random. We additionally added Gaussian noise to $\boldsymbol{x}$ to obtain the desired ISNR. For $B > 1$, reconstruction of the estimate $\widehat{\boldsymbol{x}}$ was performed using the oracle-assisted reconstruction algorithm (8) for Section 4.2 and BPDN (2) with an oracle value of $\epsilon = \|\boldsymbol{y} - \mathcal{Q}_B(\boldsymbol{y})\|_2$ for the remaining subsections. For $B = 1$, reconstruction was performed using both the *binary iterative hard thresholding* (BIHT-ℓ_1) and BIHT-ℓ_2 algorithms; the former generally performs better in lower noise scenarios and the latter performs better in higher noise scenarios [35]. We report the *reconstruction SNR* (RSNR)

$$\text{RSNR} := 10\log_{10}\left(\frac{\|\boldsymbol{x}\|_2^2}{\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2^2}\right) \tag{13}$$

in dB unless otherwise noted. Recall that the number of measurements and bit-depth are constrained by $\mathfrak{B} = MB$. We average our results over 100 trials for each parameter tuple $(N, K, \mathfrak{B}, B, \text{ISNR})$.

[4] The results of simulations did not change when the signals were DCT-sparse.



(a) ISNR = 35dB, optimal bit-depth = 8

(b) ISNR = 20dB, optimal bit-depth = 6

(c) ISNR = 10dB, optimal bit-depth = 4

(d) ISNR = 5dB, optimal bit-depth = 3

Figure 2: Oracle-assisted reconstruction error (compare to the analytical upper bound plotted in Figure 1) for $N = 1000$, $K = 10$, and $\mathfrak{B} = 3N$. As predicted by (11), the minimum reconstruction error (denoted by black dots) is achieved by smaller bit-depths as the ISNR decreases.

4.2 Oracle-assisted reconstruction

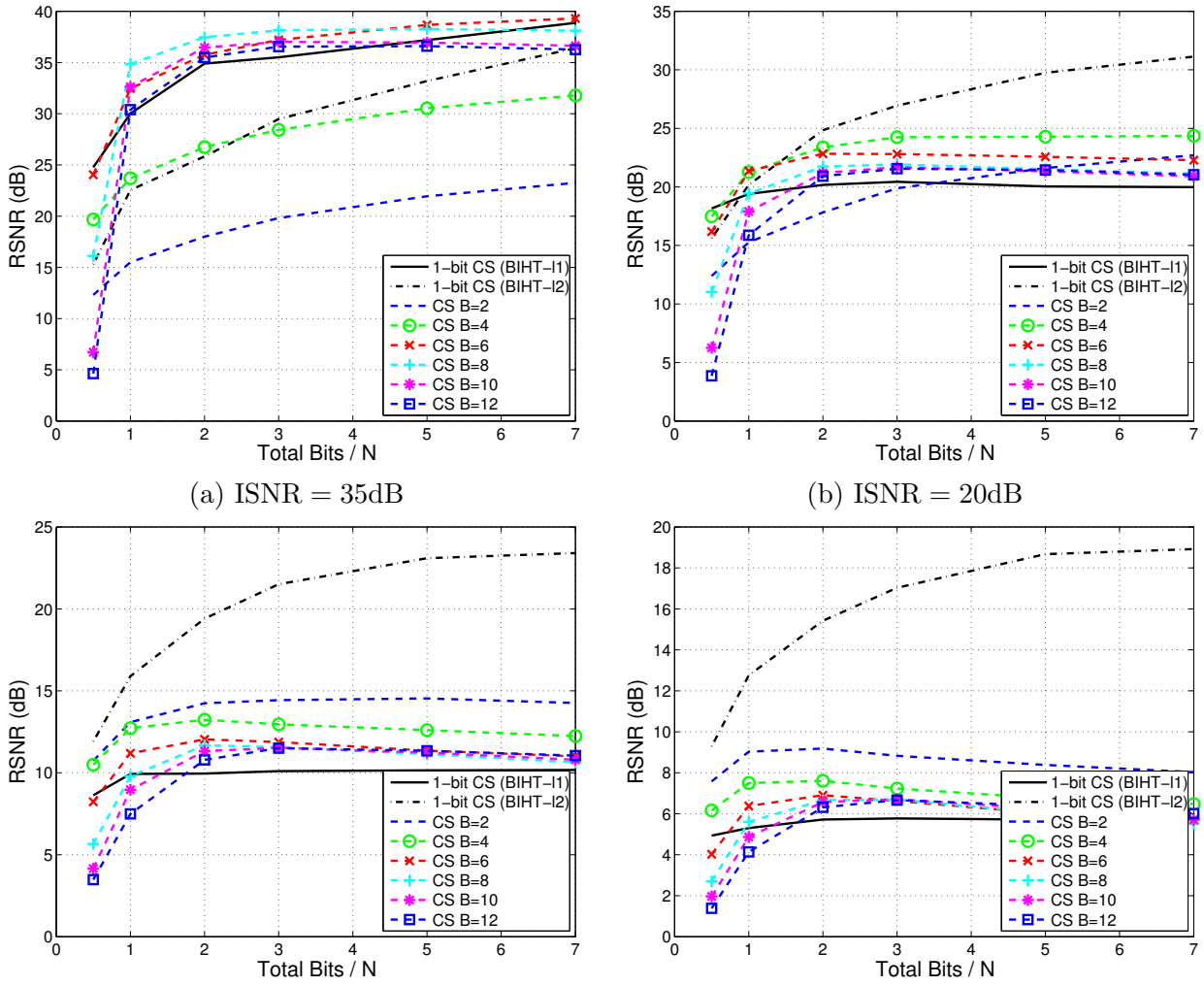
We begin by validating the message from Theorem 1, i.e., we examine the solution to the oracle-assisted reconstruction algorithm to see how the empirical performance relates to the bound (11). Our goal is to compare the performance of our simulations to the theory-based plots in Figure 1. The experiments were performed as described previously with the oracle-assisted reconstruction algorithm. We plot the reconstruction error $\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2^2$ for bit-depths between 2 and 12 for a fixed bit-budget $\mathfrak{B} = 3N$. We compared bit-depths of 2 and higher, since (11) does not hold for lower bit-depths. Furthermore, unlike the statement of Theorem 1, recall that we used a uniform quantizer and not an optimal quantizer for the Gaussian measurements. Figures 2(a)–(d) depict the results for ISNR = 35dB, 20dB, 10dB, and, 5dB, respectively.

The plots generally follow the same trends as in Figure 1; however the minimum error occurs for a slightly higher bit-depth in each case. The plots demonstrate that, as claimed in Section 3, the best performance is obtained for smaller bit-depths as the ISNR decreases.

4.3 Reconstruction performance as a function of $\mathfrak{B}$

We next explore the performance achieved using practical algorithms instead of oracle-assisted reconstruction. The experiments were performed as explained previously, for $N = 1000$ and $K = 10$, bit-depths $B = 1, 2, 4, 6, 8, 10, 12$, and for bit-budgets $\mathfrak{B} \in [N/2, 7N]$, with the BPDN and BIHT algorithms. Figures 3(a)–(d) depict the experiment for the input ISNR $= 35, 20, 10, 5$dB, respectively.

In the high ISNR regime of 35dB, bit-depths of $B = 1, 6, 8, 10$, and 12 obtain similar RSNRs of around 35dB, while smaller bit-depths result in poorer performance. This is to be expected; since when the signal noise is fairly small, we will generally do better by using more bits per measurement.



(a) ISNR = 35dB

(b) ISNR = 20dB

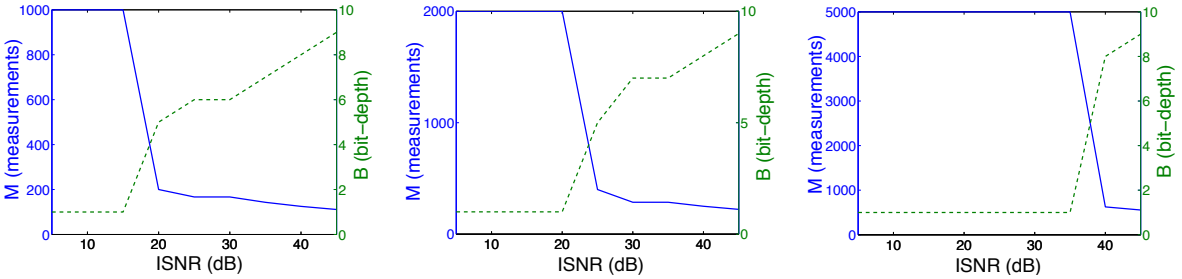
(c) ISNR = 10dB

(d) ISNR = 5dB

Figure 3: Reconstruction performance as a function of total bits, for different ISNRs. Plots depict RSNR for different bit-depths B for different ISNR with parameters $N = 1000$ and $K = 10$, and reconstruction via BPDN. The figure demonstrates that as the ISNR is decreased, smaller bit-depths achieve better performance. Additionally, 1-bit CS techniques perform competitively with or better than BPDN for all ISNRs tested.

The performance of BIHT in this case is consistent with previous results showing that the 1-bit techniques can outperform even 4-bit uniformly quantized CS measurements with BPDN recovery. This trend starts to reverse for lower signal ISNRs. Indeed for ISNRs of 10dB and 5dB, we see that 2 and 4 bit-depth quantization outperforms larger bit-depths for all budgets. Strikingly, the best performance for input SNRs of 20dB, 10dB, and 5dB is achieved by acquiring just 1 bit per measurement and reconstructing with the BIHT-ℓ_2 algorithm.

In addition to the simulations presented here, we also performed the similar simulations with $N = 1000$ and $K = 60$. We found that all of the curves in Figure 3 dropped in SNR by roughly the same constant (that depends on K). The relationship between the 1-bit curves and the others was about the same for $\mathfrak{B} = 2N$ and lower. For $\mathfrak{B} > 2N$, the 1-bit reconstructions still outperformed



(a) $\mathfrak{B} = N$ (b) $\mathfrak{B} = 2N$ (c) $\mathfrak{B} = 5N$

Figure 4: Maximum RSNR given a fixed bit-budget $\mathfrak{B}$ for parameters $N = 1000$, $K = 10$. The left side of each plot corresponds to the QC regime, while the right side corresponds to the MC regime. The solid line (blue) corresponds to the number of measurements M, while the dashed line (green) corresponds to the bit-depth B.

the others; however the performance disparity was not as great as for $K = 10$.

These simulations demonstrate two points. First, they verify that the intuition provided by the upper bound (11) is indeed correct: *for lower ISNRs it is beneficial to choose smaller bit-depths B and more measurements M.* This validates the distinction between the QC and MC regimes. Second, the 1-bit CS setup performs significantly better than the multi-bit setup for low ISNRs and is competitive with the multi-bit setup for moderate ISNRs. There are several reasons for this. When the quantization error dominates the measurement noise, the reconstruction error is primarily due to the quantization error only. This case arises when B is small; i.e., we can likely satisfy $\mathcal{Q}_B(\boldsymbol{x} + \boldsymbol{n}) = \mathcal{Q}_B(\boldsymbol{x})$ for increasing values of $|n_i|$ as B decreases. Furthermore, in this case consistent reconstruction of the 1-bit algorithms may have an advantage. Consistency could be presumably added to multibit reconstruction to improve performance but this is a topic left for future research.

4.4 Reconstruction performance as a function of ISNR

In this set of experiments, we varied the ISNR between 5dB and 45dB and searched for the (M, B) pair that maximized the RSNR, for a fixed bit-budget $\mathfrak{B}$ and parameters $N = 1000$ and $K = 10$. As demonstrated by the previous experiment, the RSNR will not be the same for each bit-budget.

Figures 4(a)–(c) depict the results of this experiment for $\mathfrak{B} = N$, $2N$, and $5N$, respectively. The left axis and solid line (blue) corresponds to the number of measurements M, while the right axis and dashed line (green) corresponds to the bit-depth B. As always, we have that $\mathfrak{B} = MB$. The QC regime is represented on the left side of the plots (low ISNR), while the MC regime is represented on the right side of the plots (high ISNR). For example, for a bit-budget of $\mathfrak{B} = 2N$, if the ISNR is 30dB, then we are operating in the MC regime and should set the bit-depth to approximately 7, resulting in the measurement ratio of approximately $M/N = 0.29$. However, for the same bit-budget, if the ISNR is 15dB, then we are operating in the QC regime and should set the bit-depth to 1, resulting in a measurement ratio of $M/N = 2$.

In each plot in Figure 4 there is a sharp transition between optimal bit-depth being high ($B \geq 5$) and low ($B \leq 2$). This transition is centered at the ISNRs 19dB, 23dB, and 38dB, for the bit-budgets $\mathfrak{B} = N$, $2N$, and $5N$, respectively. This implies that the transition occurs at higher ISNRs for higher bit-budgets. Thus, we infer that, for higher bit-budgets $\mathfrak{B}$, it is better to choose low B,

even when the input ISNR is fairly high. The bottom line then is that, for moderate ISNR, the MC regime can be assumed when the bit-budget $\mathfrak{B}$ is small, while the QC regime can be assumed when the bit-budget is large.

5 Discussion

In this paper we have studied compressive sensing (CS) systems with scalar quantization when the total number of measurement bits is fixed and noise is present on the input signal. Our results have demonstrated that *in CS, it is sometimes better to reduce the bit-depth than the number of measurements*. We found that there exist two regimes: in the high-ISNR, MC regime, we should compress by reducing the number of measurements; this regime is best suited for applications where the acquisition if each measurement is expensive. In the low-ISNR, QC regime, we should compress by reducing the number of bits per measurement; this regime is best suited to applications where the acquisition of each measurement is cheap, or large bit-depth quantizers are expensive. The key to exposing the QC regime was the recognition that there is a tradeoff between amplified input signal noise (due to the underdetermined measurement system) and the number of bits that can be allocated per measurement under a fixed bit-budget.

Choosing a low bit-depth quantizer to reduce hardware complexity while driving up the sampling rate, as is recommended for the QC regime, is not a new idea. Indeed, this same principle is the motivational force behind sigma-delta ADCs [44–47] and other non-CS oversampled ADC architectures [48–50]. However, the ideas presented here differ significantly from previous oversampled ADC architectures in the following ways: *i*) CS is compressive: Small bit-depth CS systems are expected to be used in cases where the bit-budget is significantly lower than in a conventional oversampled ADC system. The use of sparse signal models enables *compression*, i.e., a reduction in the total number of acquired bits, as opposed to just efficient sampling. *ii*) CS is non-adaptive: As described earlier, CS measurement systems are non-adaptive, meaning they do not depend on the input signal. This is true even for the 1-bit CS case. Almost all previous oversampled ADCs require some kind of feedback during quantization to produce stable representations. These differences place low bit-depth CS systems in a unique class of their own. In a few words, CS, like physics has "plenty of room at the bottom [51]."

A Proof of Theorem 1

We first extend the upper bound of Theorem 4.1 in [32] on the oracle-assisted reconstruction error to account for correlated measurement noise.

Lemma 1. *Suppose that* $\boldsymbol{y} = \Phi\boldsymbol{x}+\boldsymbol{z}$, *where* $\boldsymbol{z} \in \mathbb{R}^M$ *is a zero-mean, random vector with covariance matrix* $\Sigma = \mathbb{E}(\boldsymbol{z}\boldsymbol{z}^T)$, *and that* $\boldsymbol{x}$ *is* K*-sparse. Furthermore, suppose that* Φ *satisfies the RIP of order* K *with constant* δ. *Then the estimate* $\widehat{\boldsymbol{x}}$ *provided by the oracle-assisted reconstruction algorithm (8) satisfies*

$$\mathbb{E}\left(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2^2\right) \leq \frac{K}{1-\delta}\lambda_{\max}(\Sigma), \tag{14}$$

where $\lambda_{\max}(\Sigma)$ *is the largest eigenvalue of* Σ.

Proof. For a fixed support set $\Omega \in \{1, \ldots, N\}$ with $|\Omega| = K$, the RIP ensures that Φ_Ω is full rank, and thus the oracle estimate satisfies

$$\widehat{\boldsymbol{x}}|_\Omega = \boldsymbol{x}|_\Omega + \Phi_\Omega^\dagger \boldsymbol{z}. \tag{15}$$

We seek to estimate $\mathbb{E}\left(\|\Phi_\Omega^\dagger \boldsymbol{z}\|_2^2\right)$.

For any $K \times M$ matrix A we have that

$$\begin{aligned} \mathbb{E}\left(\|A\boldsymbol{z}\|_2^2\right) &= \mathbb{E}(\mathrm{Tr}(A\boldsymbol{z}(A\boldsymbol{z})^T)) = \mathbb{E}(\mathrm{Tr}(A\boldsymbol{z}\boldsymbol{z}^T A^T)) \\ &= \mathrm{Tr}(A\mathbb{E}(\boldsymbol{z}\boldsymbol{z}^T)A^T) = \mathrm{Tr}(A\Sigma A^T) \\ &= \sum_{j=1}^{K} \lambda_j(A\Sigma A^T), \end{aligned} \tag{16}$$

where $\lambda_j(A\Sigma A^T)$ denotes the j-th eigenvalue of $A\Sigma A^T$, and (16) follows since $A\Sigma A^T$ is a $K \times K$ matrix. Lemma 8.2 of [32] explains that the eigenvalues of this matrix can be upper bounded as

$$\begin{aligned} \lambda_{\max}(A\Sigma A^T) &\leq \lambda_{\max}(AA^T)\lambda_{\max}(\Sigma) \\ &\leq s_{\max}(A)^2 \lambda_{\max}(\Sigma), \end{aligned} \tag{17}$$

where $s_{\max}(A)$ denotes the maximum singular value of A.

Thus, to obtain the final bound, we combine (16) with (17) and substitute $A = \Phi_\Omega^\dagger$, yielding

$$\begin{aligned} \mathbb{E}\left(\|\Phi_\Omega^\dagger \boldsymbol{z}\|_2^2\right) &\leq K s_{\max}(\Phi_\Omega^\dagger)^2 \lambda_{\max}(\Sigma) \\ &\leq \frac{K}{1-\delta}\lambda_{\max}(\Sigma), \end{aligned} \tag{18}$$

since we have that $s_{\max}(\Phi_\Omega^\dagger)^2 \leq \frac{1}{1-\delta}$ from Lemma 8.1 of [32]. □

We next demonstrate that, by choosing a signal model with random values and supports, the noiseless measurements $\Phi\boldsymbol{x}$ are identically distributed and uncorrelated.

Lemma 2. *Let $\boldsymbol{x} \in \mathbb{R}^N$ be a sparse signal with support $\Omega \in \{1, \ldots, N\}$ and $|\Omega| = K$, where the elements Ω are chosen uniformly at random and the amplitudes of the non-zero coefficients are drawn according to $x_j \in \Omega \sim \mathcal{N}(0, \sigma_{\boldsymbol{x}}^2)$. Furthermore, let the $M \times N$ matrix Φ satisfy $\Phi\Phi^T = \frac{N}{M}\boldsymbol{I}_M$. Then the vector $\Phi\boldsymbol{x}$ is distributed as a mixture of Gaussians with*

$$\mathbb{E}((\Phi\boldsymbol{x})_i) = 0, \qquad \mathbb{E}((\Phi\boldsymbol{x})(\Phi\boldsymbol{x})^T) = \frac{K}{M}\sigma_{\boldsymbol{x}}^2 \boldsymbol{I}_M, \tag{19}$$

i.e., the elements $(\Phi\boldsymbol{x})_i$ of $\Phi\boldsymbol{x}$ are zero-mean uncorrelated variables.

Proof. For a fixed support Ω, each element $(\Phi\boldsymbol{x})_i$ is Gaussian distributed with mean zero since it is the sum of K zero-mean Gaussian variables. Furthermore, the distribution of $(\Phi\boldsymbol{x})_i$ over all possible supports is the sum of the distribution for each fixed support, scaled by the probability that they occur. Thus, $(\Phi\boldsymbol{x})_i$ is a mixture of Gaussians with $\mathbb{E}((\Phi\boldsymbol{x})_i) = 0$.

To derive the variance of the elements and also show that they are uncorrelated, we first examine $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T)$. The off-diagonal elements are zero, i.e., $\mathbb{E}(x_i x_j)_{i \neq j} = 0$, since the elements of $\boldsymbol{x}$ are

uncorrelated, by definition. Furthermore, the variance of the diagonal elements can be computed as

$$\mathbb{E}(x_i^2) = \sigma_{\boldsymbol{x}}^2 \mathbb{P}(i \in \Omega) = \frac{K}{N}\sigma_{\boldsymbol{x}}^2,$$

since the K non-zero support locations are chosen uniformly, any location j is chosen with probability K/N. Thus, $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T) = \frac{K}{N}\sigma_{\boldsymbol{x}}^2 \boldsymbol{I}_N$. We next compute the correlation of the measurements $\Phi\boldsymbol{x}$ to obtain

$$\begin{aligned} \mathbb{E}(\Phi\boldsymbol{x}(\Phi\boldsymbol{x})^T) &= \Phi\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T)\Phi^T \\ &= \frac{K}{N}\sigma_{\boldsymbol{x}}^2 \Phi\Phi^T = \frac{K}{M}\sigma_{\boldsymbol{x}}^2 \boldsymbol{I}_M, \end{aligned} \tag{20}$$

which concludes the proof. □

Proof of Theorem 1. Denote the error between the noiseless ideal measurements and $\boldsymbol{y}_Q$ by

$$\boldsymbol{z} := \Phi\boldsymbol{x} - \mathcal{Q}_B(\Phi\boldsymbol{x} + \Phi\boldsymbol{n}). \tag{21}$$

Our goal is to determine a bound on the variance $\sigma_{z_i}^2$ of each element z_i of $\boldsymbol{z}$. We begin by rewriting the norm squared of $\boldsymbol{z}$ as

$$\begin{aligned} z_i^2 &= [(\Phi\boldsymbol{x})_i - \mathcal{Q}_B(\Phi\boldsymbol{x} + \Phi\boldsymbol{n})_i)]^2 \\ &= [(\Phi\boldsymbol{x} + \Phi\boldsymbol{n})_i - \mathcal{Q}_B(\Phi\boldsymbol{x} + \Phi\boldsymbol{n})_i - (\Phi\boldsymbol{n})_i]^2 \\ &\leq 2[(\Phi\boldsymbol{x} + \Phi\boldsymbol{n})_i - \mathcal{Q}_B(\Phi\boldsymbol{x} + \Phi\boldsymbol{n})_i]^2 + 2(\Phi\boldsymbol{n})_i^2, \end{aligned} \tag{22}$$

where the index i denotes individual elements of the respective vector.

We now seek an upper bound on the expected value of each of the quantities in (22). We begin with the second term in (22). From the definition of Φ, we have that the elements of $\Phi\boldsymbol{n}$ have variance

$$\sigma_{\Phi\boldsymbol{n}}^2 = \mathbb{E}((\Phi\boldsymbol{n})_i^2) = \frac{N}{M}\sigma_{\boldsymbol{n}}^2, \tag{23}$$

and furthermore are uncorrelated, as was reviewed in Section 2.

To bound the first term in (22), we note that the optimal scalar quantizer of rate B for a Gaussian variable g with variance σ^2 has MSE given by $\mathbb{E}(g - \mathcal{Q}_B(g))^2 = \sigma^2 2^{-2B}$. Furthermore, the MSE of an optimal quantizer of rate B for any variable with variance σ^2 is upper bounded by that of a Gaussian variable. Our goal is to apply this quantization bound to $(\Phi\boldsymbol{x} + \Phi\boldsymbol{n})_i$. Since $(\Phi\boldsymbol{x})_i$ and $(\Phi\boldsymbol{n})_i$ are zero mean and independent of each other, then we immediately have that $\mathbb{E}\left((\Phi\boldsymbol{x} + \Phi\boldsymbol{n})_i^2\right) = \frac{K}{M}\sigma_{\boldsymbol{x}}^2 + \frac{N}{M}\sigma_{\boldsymbol{n}}^2$, where the first term follows from Lemma 2, and the second term follows from (23). Thus, we can bound the first term in (22) as

$$\begin{aligned} \mathbb{E}\left([(\Phi\boldsymbol{x} + \Phi\boldsymbol{n})_i - \mathcal{Q}_B(\Phi\boldsymbol{x} + \Phi\boldsymbol{n})_i]^2\right) &\leq \mathbb{E}\left((\Phi\boldsymbol{x} + \Phi\boldsymbol{n})_i^2\right) 2^{-2B} \\ &\leq \frac{K}{M}\sigma_{\boldsymbol{x}}^2 2^{-2B} + \frac{N}{M}\sigma_{\boldsymbol{n}}^2 2^{-2B}. \end{aligned} \tag{24}$$

Combining (23) and (24) as in (22) yields

$$\sigma_{z_i}^2 \leq 2\frac{K}{M}\sigma_{\boldsymbol{x}}^2 2^{-2B} + 2\frac{N}{M}\sigma_{\boldsymbol{n}}^2\left(1 + 2^{-2B}\right). \tag{25}$$

We have thus far established an upper bound on the variance $\sigma_{z_i}^2$ of the error z_i of each measurement. We next obtain a bound on the eigenvalues of the covariance matrix $\Sigma = \mathbb{E}(\boldsymbol{z}\boldsymbol{z}^T)$. The off-diagonal elements of Σ can be written as

$$\begin{aligned}\mathbb{E}(z_i z_j)_{i\neq j} &= \mathbb{E}((\Phi\boldsymbol{x})_i(\Phi\boldsymbol{x})_j)) - \mathbb{E}((\Phi\boldsymbol{x})_i\mathcal{Q}_B(\Phi\boldsymbol{x}+\Phi\boldsymbol{n})_j) \\ &\quad -\mathbb{E}((\Phi\boldsymbol{x})_j\mathcal{Q}_B(\Phi\boldsymbol{x}+\Phi\boldsymbol{n})_i) + \mathbb{E}(\mathcal{Q}_B(\Phi\boldsymbol{x}+\Phi\boldsymbol{n})_i\mathcal{Q}_B(\Phi\boldsymbol{x}+\Phi\boldsymbol{n})_j) \\ &= -\mathbb{E}(\mathcal{Q}_B(\Phi\boldsymbol{x}+\Phi\boldsymbol{n})_i\mathcal{Q}_B(\Phi\boldsymbol{x}+\Phi\boldsymbol{n})_j), \end{aligned} \tag{26}$$

since $\mathbb{E}((\Phi\boldsymbol{x})_i(\Phi\boldsymbol{x})_j)) = 0$ by design and, for an optimal scalar quantizer, we have that $\mathbb{E}(\mathcal{Q}_B(\Phi\boldsymbol{x}+\Phi\boldsymbol{n})_i\mathcal{Q}_B(\Phi\boldsymbol{x}+\Phi\boldsymbol{n})_j) = \mathbb{E}((\Phi\boldsymbol{x})_j\mathcal{Q}_B(\Phi\boldsymbol{x}+\Phi\boldsymbol{n})_i)$ [43]. Thus, the matrix Σ has $\sigma_{z_i}^2$ along its diagonal and $\mathfrak{S}$ for all other entries. We next apply Gershgorin's circle theorem, which explains that any eigenvalue is upper bounded by the diagonal entry plus the sum of the magnitudes of the off-diagonal entries of each row of Σ. Thus, we have

$$\lambda_{\max}(\Sigma) \leq \sigma_{z_i}^2 + (M-1)\mathfrak{S}, \tag{27}$$

where $\mathfrak{S} = \max_{i\neq j} |\mathbb{E}(z_i z_j)|$.

To obtain the final bound, we combine to (25) with (27) and apply the upper bound in Lemma 1. We express the bound with the substitution $M = \mathfrak{B}/B$. □

Acknowledgments

Thanks to Petros Boufounos, Mark Davenport, Vivek Goyal, Laurent Jacques, Christoff Studer, and John Treichler for useful discussions and sage advice.

References

[1] E. Candès, "Compressive sampling," in *Proc. Int. Congress Math.*, Madrid, Spain, Aug. 2006.

[2] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 6, no. 4, pp. 1289–1306, 2006.

[3] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[4] J. Tropp and A. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[5] J. Tropp, J. Laska, M. Duarte, J. Romberg, and R. Baraniuk, "Beyond Nyquist: Efficient sampling of sparse, bandlimited signals," *IEEE Trans. Inform. Theory*, vol. 56, no. 1, pp. 520–544, 2010.

[6] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Mag.*, vol. 25, no. 2, pp. 83–91, 2008.

[7] J. P. Slavinsky, J. Laska, M. Davenport, and R. Baraniuk, "The compressive mutliplexer for multi-channel compressive sensing," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[8] W. Bajwa, J. Haupt, G. Raz, S. Wright, and R. Nowak, "Toeplitz-structured compressed sensing matrices," in *Proc. 14th IEEE/SP Workshop Statistical Signal Processing (SSP'07)*, Madison, WI, Aug. 2007.

[9] M. Duarte and R. Baraniuk, "Spectral compressive sensing," 2010, Preprint.

[10] C. Hegde, M. Duarte, and V. Cevher, "Compressive sensing recovery of spike trains using a structured sparsity model," in *Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, Saint-Malo, France, Apr. 2009.

[11] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inform. Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[12] E. Hale, W. Yin, and Y. Zhang, "A fixed-point continuation method for ℓ_1-regularized minimization with applications to compressed sensing," Rice Univ., CAAM Dept., Tech. Rep. TR07-07, 2007.

[13] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing," *SIAM J. Imag. Sci.*, vol. 1, no. 1, pp. 143–168, 2008.

[14] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projections for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Select. Top. Signal Processing*, vol. 1, no. 4, pp. 586–597, 2007.

[15] E. van den Berg and M. Friedlander, "Probing the Pareto frontier for basis pursuit solutions," *SIAM J. on Sci. Comp.*, vol. 31, no. 2, pp. 890–912, 2008.

[16] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, 2009.

[17] T. Blumensath and M. Davies, "Iterative hard thresholding for compressive sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, 2009.

[18] D. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proc. Natl. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[19] M. Lustig, D. Donoho, and J. Pauly, "Rapid MR imaging with compressed sensing and randomly under-sampled 3DFT trajectories," in *Proc. Annual Meeting of ISMRM*, Seattle, WA, May 2006.

[20] D. Healy. (2005) Analog-to-information. DARPA BAA #05-35. [Online]. Available: http://www.darpa.mil/mto/solicitations/baa05-35/s/index.html

[21] C. S. Gunturk, A. Powell, R. Saab, and O. Yilmaz, "Sobolev duals for random frames and sigma-delta quantization of compressed sensing measurements," 2010, preprint.

[22] P. Boufounos, "Universal rate-efficient scalar quantization," 2010, preprint.

[23] C. S. Gunturk, A. Powell, R. Saab, and O. Yilmaz, "Sobolev Duals for Random Frames and Sigma-Delta Quantization of Compressed Sensing Measurements," 2010, preprint.

[24] A. Zymnis, S. Boyd, and E. Candès, "Compressed sensing with quantized measurements," *IEEE Sig. Proc. Letters*, vol. 17, no. 2, Feb. 2010.

[25] J. Sun and V. Goyal, "Quantization for compressed sensing reconstruction," in *Proc. Sampling Theory and Applications (SampTA)*, Marseille, France, May 2009.

[26] J. Z. Sun and V. K. Goyal, "Optimal quantization of random measurements in compressed sensing," in *Int. Sym. on Inform. Theory (ISIT)*, June 2009.

[27] H. Q. Nguyen, V. Goyal, and L. Varshney, "Frame permutation quantization," *Appl. Comput. Harmon. Anal.*, Nov. 2010.

[28] J. Laska, P. Boufounos, M. Davenport, and R. Baraniuk, "Democracy in action: Quantization, saturation, and compressive sensing," *to appear in App. Comp. and Harm. Anal.*, 2011.

[29] L. Jacques, D. Hammond, and M. Fadili, "Dequantizing compressed sensing: When oversampling and non-gaussian contraints combine," *IEEE Trans. Inform. Theory*, vol. 57, no. 1, pp. 559–571, 2009.

[30] E. Candès and M. A. Davenport, "How well can we estimate a sparse vector?" 2011, preprint.

[31] J. Treichler, M. Davenport, and R. Baraniuk, "Application of compressive sensing to the design of wideband signal acquisition receivers," in *U.S./Australia Joint Work. Defense Apps. of Signal Processing (DASP)*, Lihue, Hawaii, Sept. 2009.

[32] M. Davenport, J. Laska, J. Treichler, and R. Baraniuk, "The pros and cons of compressive sensing: Noise folding and dynamic range," 2011, preprint.

[33] T. Cover and J. Thomas, *Elements of Information Theory.* New York, NY: Wiley-Interscience, 1991.

[34] S. Sarvotham, D. Baron, and R. Baraniuk, "Measurements vs. bits: Compressed sensing meets information theory," in *Proc. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Sept. 2006.

[35] L. Jacques, J. Laska, P. Boufounos, and R. Baraniuk, "Robust 1-bit compressive sensing via binary ϵ-stable embeddings," 2011, preprint.

[36] B. Le, T. W. Rondeau, J. H. Reed, and C. W. Bostian, "Analog-to-digital converters," *IEEE Sig. Proc. Mag.*, Nov. 2005.

[37] P. Boufounos and R. Baraniuk, "1-bit compressive sensing," in *Proc. Conf. Inform. Science and Systems (CISS)*, Princeton, NJ, Mar. 2008.

[38] J. Laska, Z. Wen, W. Yin, and R. Baraniuk, "Trust, but verify: Fast and accurate signal recovery from 1-bit compressive measurements," *IEEE Trans. Sig. Proc.*, vol. 59, no. 11, pp. 5289–5301, 2011.

[39] P. Boufounos, "Greedy sparse signal reconstruction from sign measurements," in *Proc. Asilomar Conf. on Signals Systems and Comput.*, Asilomar, California, Nov. 2009.

[40] E. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Stat.*, vol. 35, no. 6, pp. 2313–2351, 2007.

[41] Y. Plan and R. Vershynin, "One-bit compressed sensing by linear programming," 2011, preprint.

[42] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[43] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Info. Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[44] J. C. Candy and G. C. Temes, Eds., *Oversampling Delta-Sigma Converters.* IEEE Press, 1992.

[45] P. M. Aziz, H. V. Sorensen, and J. V. D. Spiegel, "An overview of Sigma-Delta converters," *IEEE Sig. Proc. Mag.*, vol. 13, no. 1, pp. 61–84, Jan. 1996.

[46] J. J. Benedetto, A. M. Powell, and O. Yilmaz, "Sigma-delta quantization and finite frames," *IEEE Trans. Info. Theory*, vol. 52, no. 5, pp. 1990–2005, May 2006.

[47] P. Boufounos, "Quantization and Erasures in Frame Representations," Ph.D. dissertation, MIT EECS, Cambridge, MA, Jan. 2006.

[48] Z. Cvetković and I. Daubechies, "Single-bit oversampled A/D conversion with exponential accurace in the bit-rate," *IEEE Trans. Info. Theory*, vol. 53, no. 11, pp. 3979–3989, 2007.

[49] V. K. Goyal, M. Vetterli, and N. Thao, "Quantized overcomplete expansions in $\mathbb{R}^n$: Analysis, synthesis, and algorithms," *IEEE Trans. Inform. Theory*, vol. 44, no. 1, pp. 16–31, 1998.

[50] S. Hoyos, B. Sadler, and G. Arce, "Monobit digital receivers for ultrawideband communications," *IEEE Trans. Wireless Comm.*, vol. 4, no. 4, pp. 1337–1344, July 2005.

[51] R. Feynman, "There's plent of room at the bottom," in *Caltech Eng. and Sci.*, American Physical Society, Ed., vol. 23, no. 5, 1960, pp. 22–36.